DOCUMENT RESUME

ED 328 577 TM 016 082

AUTHOR Ligon, Glynn; And Others
TITLE Statewide Testing in Texas. A Symposium Presented at the Annual Meeting of the Southwest Educational Research Association (Austin, Texas, January 25-27, 1990).
INSTITUTION Southwest Educational Research Association.
PUB DATE Jan 90
NOTE 24p.
PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Trends; Elementary Secondary Education; *History; State Government; *State Programs; State School District Relationship; Test Construction; *Testing Programs
IDENTIFIERS Test Publishers; *Texas

ABSTRACT

Four essays and associated introductory text constituting the proceedings of a symposium on statewide testing in Texas are presented. The papers include: (1) "Statewide Testing in Texas: Historical Perspective" (Glynn Ligon); (2) "Statewide Testing in Texas: Local School District Perspectives" (Whit Johnstone and others); (3) "Statewide Testing in Texas: Test Author/Publisher Perspectives" (Mikel Brightman and others); and (4) "Statewide Testing in Texas: Governmental Perspectives" (Elaine Davis). Audience questions and comments are included. Recommendations and comments are made concerning the following topics: defining who is being held accountable and how; the three most important issues in statewide testing programs (money, time, and usefulness of data); the need for more focus on the sample of items used for norm-referenced scores on subsequent statewide tests; problems associated with legislators' role in educational policymaking; the need for interlevel articulation of the reasons for testing and how research will influence testing programs; and related testing problems. It is concluded that: Texas schools have become more accountable; education in Texas has improved for lower achieving students, but not for higher achieving students. (TJH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED328577

# Statewide Testing in Texas

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

GLYNN LIGON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

*A Symposium*
*Presented at the*
*Annual Meeting of the*
*Southwest Educational Research Association*

*January 27, 1989*
*Houston, Texas*

TM016082

# *Statewide Testing in Texas*

Dr. Glynn Ligon, Chair
Austin ISD
Dr. Whit Johnstone
Irving ISD
Dr. Evangelina Mangino
Austin ISD
Dr. Carl Shaw
Houston ISD
Dr. Mikel Brightman
The Psychological Corporation
Dr. Paul Williams
CTB/McGraw-Hill
Dr. H.D. Hoover
University of Iowa
Dr. Elaine Davis
Office of the Governor of Texas


***A Symposium Presented at the Annual Meeting of the Southwest Educational Research Association***

*January 27, 1989*
*Houston, Texas*


This monograph is a 1989 publication of the Southwest Educational Research Association.

# *Statewide Testing in Texas*

## *CONTENTS*


| Section | Page |
|---|---|
| Foreword | i |
| Definitions | ii |
| Executive Summary | iii |
| Participants | v |
| ***Historical Perspective:*** | |
| Glynn Ligon, Chair; Austin ISD | 1 |
| ***Local School DistrictPerspectives:*** | |
| Whit Johnstone, Irving ISD | 3 |
| Evangelina Mangino, Austin ISD | 5 |
| Carl Shaw, Houston ISD | 7 |
| ***Test Author/Publisher Perspectives:*** | |
| Mikel Brightman, The Psychological Corporation | 8 |
| Paul Williams, CTB/McGraw-Hill | 10 |
| H.D. Hoover, University of Iowa | 12 |
| ***Governmental Perspectives:*** | |
| Elaine Davis, Office of the Governor of Texas | 14 |
| ***Summary Comments*** | 17 |

# *Statewide Testing in Texas*

## *FOREWORD*

*Glynn Ligon*

**This month, Texas completes its ninth year of statewide testing,** and in May/June will award diplomas to its third class of graduates who were required to pass competency tests in both mathematics and language arts. As much as some have protested the influence of the current testing program, others are calling for expansion of testing to more and higher skill areas. Behind the public controversies brews a myriad of technical, psychometric issues that challenge the reliability, even the validity, of the statewide testing program.

Students should be required to demonstrate basic competencies before receiving credit for passing their basic courses rather than earning spurious credit only to be exposed for illiteracy or mathematical inability by an add-on examination. However, in the absence of an unexpected, collective insight in Texas and other states, the Texas Legislature, Texas Education Agency, and the State Board of Education must make some immediate decisions about the direction of the statewide testing program.

To contribute to the dialogue that will lead to the making of these critical decisions, the Southwest Educational Research Association sponsored a symposium on *Statewide Testing in Texas* at its annual meeting on January 27, 1989, in Houston. This document is a summary of the comments, suggestions, and challenges offered by the distinguished participants in that symposium.

# *Statewide Testing in Texas*

## *DEFINITIONS*

| | |
|---|---|
| CRT | Criterion-referenced test: A CRT measures mastery of specific objectives. The TEAMS is a CRT. |
| Imbedded Items | Mixing items from an NRT among CRT items to obtain both national percentile ranks and mastery scores for students. |
| NAEP | National Assessment of Educational Progress: A national achievement testing program. |
| Norms | Comparison scores from a nationwide sample of students. |
| NRT | Norm-referenced test: An NRT measures a broad range of skills and ranks a student in relation to a national sample of students. |
| SBOE | State Board of Education |
| TABS | Texas Assessment of Basic Skills: Texas' statewide test from 1981 to 1984. |
| TAP | Texas Assessment Program: Texas' first statewide testing program. |
| TEA | Texas Education Agency |
| TEAMS | Texas Educational Assessment of Minimum Skills: Texas' statewide test since 1985. |

# *Statewide Testing in Texas*

## *EXECUTIVE SUMMARY*

**As the Texas Legislature in Austin** began to deliberate the future of statewide testing in Texas, a group of involved professionals met in Houston at the Annual Meeting of the Southwest Educational Research Association to define and explore current issues and opportunities in statewide testing. Three local school district testing administrators and three representatives of test publishers presented historical views and raised substantive issues about testing programs, followed by comments from a representative of the Governor's Office.

The audience formed a clear impression that statewide achievement testing programs raise a myriad of issues— psychometric, financial, instructional, political, and practical. There was not unanimity of opinion on these issues among the experts in this symposium. In summary, the major issues that must be considered by the Legislature, and reconsidered by the Texas Education Agency and the State Board of Education are:

1. **October Testing Dates** — Will we sacrifice accountability without gaining timely return of test results?

2. **Imbedding or Appending NRT Items** — Will we sacrifice the quality of the current CRT by mixing in a too small number of NRT items to yield reliable national norms?

3. **Cost** — Will we continue to pay for a State test that adds to the testing burden rather than replacing local testing programs?

4. **Ranking of Schools** — Will we continue to rank schools by reducing all test results to a single score that purports to compare equitably across elementary and high schools and across three subject areas?

5. **Accountability** — Are Texas schools more accountable now after eight years of statewide testing?

6. **Quality of Education** — Are Texas students better educated now after eight years of statewide testing?

***Elected officials, educators, parents, and other taxpayers*** *are encouraged to read this monograph as information for forming their own opinions on the future direction for statewide testing in Texas.*

# *Statewide Testing in Texas*

## *PARTICIPANTS*

Mikel Brightman, Ph.D.
*Senior Project Director, Educational Measurement Division, The Psychological Corporation*

Elaine Davis, Ph.D.
*Policy Analyst, Governor's Office of Budget and Planning*

H.D. Hoover, Ph.D.
*Professor, University of Iowa; Director, Iowa Basic Skills Testing Program*

Whit Johnstone, Ph.D.
*Director of Research and Planning, Irving ISD*

Glynn Ligon, Ph.D.; Chair
*Executive Director, Department of Management Information, Austin ISD*

Evangelina Mangino, Ph.D.
*Evaluator, Systemwide Testing, Office of Research and Evaluation, Austin ISD*

Carl Shaw, Ph.D.
*Director of Test Development, Houston ISD*

Paul Williams, Ph.D.
*Director of Research and Measurement, CTB/McGraw-Hill*

# *Statewide Testing in Texas*

***-- Historical Perspective***

***Glynn Ligon***
***Austin ISD***

**Statewide testing in Texas** has been somewhat like the story of the motorist who ran out of gas and walked up to a rancher's house for assistance. On the way to the barn to get some gas, the motorist was compelled to ask about all the targets posted on the rancher's trees and the arrows shot squarely in the bull's eye of each one. The rancher laughed and informed the motorist that his daughter got a bow and arrow set for Christmas and just goes around shooting arrows. If one happens to hit a tree, then she goes over and sticks a target over the arrow.

*"Texas' statewide testing program has been controversial from the start and remains so."*

**1980** - At this time everyone was talking mastery of basic and minimum skills. Texas jumped into this area a little bit late. The first statewide testing was the Texas Assessment of Basic Skills (TABS) for grades 3, 5, and 9. The purpose of TABS was to provide a tool for state and local educators to identify where help was needed. The Texas Education Agency (TEA) encouraged the use of TABS for diagnostics and local interpretation of needs. TABS lasted for five administrations.

**1985** - Texas Educational Assessment of Minimum Skills (TEAMS) was begun for grades 1, 3, 5, 7, 9, and 11. The nature of the TABS did not change much, but the use of the test changed. Thanks to the Accreditation Division of TEA and heavy media publicity, TEAMS began to be used to rank individual schools and districts. Rankings were published in newspapers across the State.

*"Thanks to the Accreditation Division of TEA and heavy media publicity, TEAMS began to be used to rank individual schools and districts."*

**1990** - The third round of statewide testing will begin. The test will be expanded to give a nationally norm- referenced percentile, and the date of the testing will be moved from spring to October. The change in date is designed to release teachers from accountability/responsibility for test scores and move back to the purpose of using the test for diagnostics.

Texas' statewide testing program has been controversial from the start and remains so. Some of these controversies have been:

1. Statewide testing is perceived as an intrusion into traditional local control of schools in Texas.
2. Instructional time is being diverted from the established curriculum to testing.
3. Control of the curriculum is shifting to the State with an over-emphasis on just the basic skills.

# *Statewide Testing in Texas*

*-- Historical Perspective*

...continued

*Glynn Ligon*

4. Accountability and evaluation of individual teachers is being based upon TEAMS scores.

5. Campus-level reporting of test results has led to ranking of campuses and districts.

7. The cost for TEAMS is high during a period of tight budgets.

8. The Written Composition test has been unreliable, using a controversial scoring technique which has failed a disproportionate number of gifted students.

9. The difficulty of individual objectives has drifted from year to year making comparisons inappropriate.

10. The State Board of Education's seventy percent mastery criterion has been translated into percentages of items correct that range from 61% to 89% across test levels.

11. Scale scores were developed for use in keeping the tests of equal difficulty from year to year, but have been used to compare scores inappropriately across grades and test areas.

12. WAVE scores have been developed to combine TEAMS scores across grades and areas to obtain a single score for entire schools and for entire districts.

13. TEA recommended and the State Board of Education approved a contract for imbedding (or appending) of NRT items into the TEAMS, thus creating a longer test with questionable normative reliability.

Now we face the issue of combining 20 norm-referenced items with a criterion-referenced test. The State Board of Education, on the recommendation of TEA, issued a request for proposals to test publishers for the third five-year cycle of tests and called for inclusion of norm-referenced items to provide a reliable percentile rank score for each student tested. Can this be done without making the test too long. Are there other options that should be considered? Our experts were provided lists of questions and issues to be discussed.

*"The cost for TEAMS is high during a period of tight budgets."*

*"TEA recommended and the State Board of Education approved a contract for imbedding (or appending) of NRT items into the TEAMS, thus creating a longer test with questionable normative reliability."*

# *Statewide Testing in Texas* -- *Local School District Perspectives*

***Whit Johnstone***
*Irving ISD*

**Before there was TABS** or TEAMS, the TAP program was a NAEP-like assessment, given to randomly selected samples of students across Texas. It was a good plan for statewide assessment and provided statewide statistics. It did not provide data for individual districts. Then the law was passed which established the TABS to provide data for individual districts. The TABS focused on areas of curriculum covered by all districts in Texas.

Local testing programs vary considerably. Most local programs are norm-referenced and use commercial nationally normed tests with grade-level norms. Sometimes an achievement test is used at the local school level. Irving ISD tests every year at every grade level. Such testing provides comparable information from one grade level to the next that is current and up-to-date.

Nationally normed achievement tests contain common threads of curriculum from across the country. By using this type of test, educators can tell when there are significant gaps in the local curriculum as compared to the nation. Information is provided on how local students compare to the average student in the nation. This type of test also provides a validity check of the local program compared to the nation.

*(Glynn Ligon: Would you consider replacing the NRT given in Irving with the statewide test if it were given in October?)*

I would consider replacing Irving's NRT with the statewide test, depending on how well the information from the State program could be substituted for the data provided by the current local NRT.

What use is the Texas statewide testing program to local school districts? Written into the legislation which created the TABS were rules and interpretations from the State Board of Education. Although districts had little input into the rules, clear direction was provided for what to do with the results. Results were to be used to diagnose instructional strengths and weaknesses for individual students to place them in remedial programs. At the campus level, results were to be used to develop plans to improve instructional programs. Results were to be reported to local school boards and to the press by campus.

Districts responded in different ways to the TABS. Some districts had excellent criterion-referenced testing programs that they threw out when statewide testing rules came up. Some districts continued their local testing programs.

*"Some districts had excellent criterion-referenced testing programs that they threw out when statewide testing rules came up."*

# *Statewide Testing in Texas* -- *Local School District Perspectives*

...continued

***Whit Johnstone***

The TEAMS was a great leveler for assessment in Texas. Districts had been in different places with their testing programs. The TEAMS has not had a great impact on diagnosing strengths and weaknesses. The actual impact of the TEAMS on districts has been through the press which reported TEAMS scores and used the results to compare schools and especially districts. There was appropriate interest by the media in the quality of schools. The press led the State in this area. The reforms of House Bill 72 followed, and the statewide test changed from the TABS to the TEAMS. There were new requirements for comparing districts and norm-referenced data were provided to the Legislature.

As a result of what the media were doing, in self-defense, districts designed programs to remediate all students, even students who had no problems mastering the TEAMS. Schools and districts looked better when students achieved mastery with four-out-of-four items measured, instead of three-out-of-four items. Districts started competing for media attention, money, staff, and support. Districts were assigned a grade through TEAMS. The grade became important. TEA got pulled into this comparison across districts, and rankings of districts followed.

*"As a result of what the media were doing, in self-defense, districts designed programs to remediate all students, even students who had no problems mastering the TEAMS."*

# *Statewide Testing in Texas* -- *Local School District Perspectives*

***Evangelina Mangino***
***Austin ISD***

**The three most important practical issues** in testing are:

- Money,
- Time, and
- Usefulness of data.

The State testing program mandates testing of all students in grades 1, 3, 5, 7, 9, and 11 in all public schools, including those who already have local testing programs. In districts with local testing programs covering all grades, half of the students must be tested at least twice a year with achievement tests. If each test, on the average, takes four hours (testing and handing out materials) and 70% of the students in the state are tested with norm-referenced tests, there are approximately 1,000,000 students who spent at least 8 hours on achievement tests in 1987-88.

The TEAMS, as its name indicates, is designed to measure only minimum skills. Therefore, the TEAMS is a test with a very low ceiling. NRT's yield a broader picture, which allows the districts to evaluate achievement gains for students at all levels of achievement. In addition, there is the need for NRT results for federal program evaluations.

Plans by TEA to expand the State testing program to include national percentiles and rotating objectives to cover all the "testable" essential elements are encouraging. If these plans are carried out adequately, districts would be able to substitute the use of NRT's with the Texas state test in the odd-numbered grades.

School districts currently using NRT's pay for testing materials and related expenses out of their local budgets. Substituting the use of NRT's with the State-mandated test would cut testing cost in half for these districts. Although these savings would be attractive to districts, this budgetary advantage is offset by the technical disadvantage of not having comparable data from year to year (the State test would be administered at odd-numbered grades in October while most local NRT's are given in the spring).

According to the SBOE request for proposals for the 1990-1995 Texas testing program, the new test will yield a national percentile based on the test selected. TEA selected the Stanford, published by the Psychological Corporation. This test will be used as a base to customize a test for Texas which will cover the essential elements and yield a Stanford-equivalent national percentile.

Experiences in customizing tests (such as the MAT-6 for New York City) have shown that the norm data obtained by those

*"The three most important practical issues in testing are:*
- *Money,*
- *Time, and*
- *Usefulness of data."*

*"Experiences in customizing tests have shown that the norm data obtained by those tests are questionable/unreliable."*

# *Statewide Testing in Texas* -- *Local School District Perspectives*

...continued

***Evangelina Mangino***

tests are questionable/unreliable.

The preference of the Texas urban districts, and no doubt, of many other districts in the State, is that an NRT be adopted statewide and that the State continue to pay (with the $3.40 per student currently withheld from districts' State Compensatory Education funds) for the odd-numbered grades. Items covering the essential elements but not covered by the NRT would have to be administered as a supplementary CRT. This combination of NRT and supplementary items would assure the measurement of mastery of essential elements, and it would yield true norm data.

One condition that must be met in order for districts to replace their current NRT's at even-numbered grades with the NRT selected by the State, would be that the State test be administered to odd-numbered grades at the same time as districts administer the test to even-numbered grades. The Big Eight districts in Texas test in the spring. Unfortunately, at this time, the proposed date for the 1990-95 testing program is October for all grades. The reason TEA gives for October testing is to collect data at the beginning of the year to give teachers diagnostic information and to reduce the anxiety of teachers about accountability. Currently, scoring of the TEAMS at grades 3, 5, 7, and 9 takes two and a half months. If students were tested in October, results would not be available to teachers and administrators until the end of December or beginning of January. These data would hardly be useful for planning at the beginning of the year.

Removing pressure from the teachers also reduces accountability and the sense of ownership. Testing at the beginning of the school year would be a somewhat useless exercise to most districts, especially those which lack the computer and programming facilities necessary to analyze the results based on previous school-year attendance information by campus.

Spring testing yields summative information for the year and allows three-month-old data to be used for planning at the beginning of a new school year. Testing in October would result in using eleven-month old data if the test results are to be used for planning at the beginning of the next year.

In conclusion, we would like to see a shelf norm-referenced test adopted with a supplementary CRT to cover additional essential elements. We would like to administer these tests in the spring so that we could combine it with our local programs and save at least half of the money currently used for duplicate testing (which would, according to our estimates, be about $8,000,000 to districts). Spring data would be more timely and useful for planning and accountability.

*"If students were tested in October, results would not be available to teachers and administrators until the end of December or beginning of January."*

*"The preference of the Texas urban districts and many others is that an NRT be adopted statewide, and that the State continue to pay for odd-numbered grades.*

*... we would like to see a shelf norm-referenced test adopted, with a supplementary CRT to cover additional essential elements."*

# *Statewide Testing in Texas* -- *Local School District Perspectives*

***Carl Shaw***
***Houston ISD***

**I prefer to refer to time lost** to testing rather than time *devoted* to testing. In Houston there are custom-built testing programs, so the state tests are not the only show in town. There are 14 subdistricts in Houston that also have their own testing programs going. When you add them all up, the district loses at least eight hours to actual test taking, but much more to the preparation and adjustments to schedules associated with testing. It is hard to get a handle on time lost just to TEAMS. We need to take into account instruction that is solely for test preparation.

*(Glynn Ligon: The official TEA position is that time devoted to testing is academic engaged time.)*

In Houston, the superintendent is very concerned about the outcome of TEAMS testing. Then the parents are concerned about their kids being pulled out of regular classroom instruction for instruction on TEAMS. It was the purpose, the goal of TEAMS that items/skills be subsumed in the curriculum, not that the TEAMS should assume the curriculum.

There is also the issue of cheating, which is stealing from the students. Examples:

- We cheat our students out of class work when we pull them from class for unnecessary TEAMS preparation.
- I am sure that there is a Xerox copy of the test in the files somewhere in HISD.
- Smiles or frowns from the teacher looking at students' answers during the test are also cheating.
- Sometimes time limits are stretched.

The higher the impact of the program, the more likely cheating is going on. In HISD, teachers have lost jobs helping kids cheat. Cheating may initially raise scores, but after three to four years, I wonder if it will not lower scores.

*(Glynn Ligon: TEA and the State Board are concerned about cheating. That is one impetus for the proposed move to October testing and wider objectives.)*

With October testing, you will not have new classroom instruction start until November, and you will not have results back to the classroom teachers until December at the earliest.

*"I prefer to refer to time lost to testing rather than time devoted to testing. ...parents are concerned about their kids being pulled out of regular classroom instruction for instruction on TEAMS."*

*"The higher the impact of the program, the more likely cheating is going on. Smiles or frowns from the teacher looking at students' answers during the test are... cheating."*

*"With October testing, you will not have new instruction until November, and you will not have results back to the classroom teachers until December at the earliest."*

# *Statewide Testing in Texas* -- *Test Publisher Perspectives*

***Mikel Brightman***
***The Psychological Corporation***

**Currently only one state** does not have a statewide testing program.

| Year of Implementation | States with Statewide Programs |
|---|---|
| Pre-1970 | 8 |
| 1970-79 | +28 |
| 1980-88 | +14 |
| | TOTAL: 50 |
| NAEP(1990) | 20 |

*"Currently only one state does not have a statewide testing program."*

Background and contextual issues need to be kept in mind when interpreting assessment program results.

- State budgets
- 5-7% of GNP for education
- Dropout rates are increasing
- Per pupil costs are increasing
- SAT scores are declining
- Lower pupil/teacher ratios

Different test types are being used.

| Test Types | Number of States |
|---|---|
| NRT | 30 |
| CRT/Custom | 33 |
| Both | 14 |
| None | 1 |

Where does the responsibility reside for these statewide programs?

| Responsible Agency | Number of States |
|---|---|
| State | 35 |
| Local Districts | 2 |
| State/Local Districts | 9 |
| None | 5 |

*"In the past the accountability lay with the student, the mom, and the teacher."*

In the past the accountability lay with the student, the mom, and the teacher.

# *Statewide Testing in Texas* -- *Test Publisher Perspectives*

...continued

***Mikel Brightman***

Reporting responsibility is different from state to state.

| Agency Responsible for Reporting | Number of States |
|---|---|
| Individual School | 25 |
| District | 37 |
| State | 43 |
| States Allowing Comparisons | 38 |
| States Reporting Demographics | 21 |
| States with Rewards/Sanctions | 25 |

There is a concern with the quality of education for certain classes of students. The Psychological Corporation (for TEA) is building a new statewide testing program, implementing new technology. There is currently no agreement whether norm-referenced items will be imbedded or appended. Pilot studies are planned for the fall. There will be 20 items per content domain with a balance between reliability (based on the number of items) and time required by the test.

The MAT-6/TEAMS equating technique was poor because the TEAMS matched with only some of the MAT-6 items/content. The distribution of TEAMS test scores, their limited variance, prevented a good equating. The new contract takes a completely different approach.

The goal is to get a test that is psychometrically defensible and curricularly defensible.

*"There is currently no agreement whether norm- referenced items will be imbedded or appended.*

*The goal is to get a test that is psychometrically and curricularly defensible."*

# *Statewide Testing in Texas* -- *Test Publisher Perspectives*

***Paul Williams***
***CTB/McGraw-Hill***

**I urge caution** in rushing to innovate. California put a burden on local school systems with graduate competency programs that are injurious to students. The standards change from year to year. There is a rush to innovate ahead of our technology.

In the early 70's norm-referenced tests were incorporated into school testing programs on a regular basis, and objective reporting of group changes were possible. These tests reported the students' progress from year to year and against national norms. The current movement, however, is a movement towards customized testing.

Districts and/or states want custom tests, so that they can judge how well their students are learning the specific content they want taught. But they also want to compare to national norms, so they want two tests, or they want to somehow combine criterion-referenced items and norm-referenced items. But they do not want to spend all their time testing, or all their money on testing programs, so they want to somehow combine the two types into a customized test and use both types of reporting.

A major issue is when are norms valid and when are they not valid? We have taken shelf tests and manipulated them to be customized tests. The problem is that they are then not norm valid.

If we have a nonrandom sample of content, students may be tested on only fractions, where the original normed set of questions included both fractions and decimals. Then if teachers emphasize fractions, test scores will go up, and the impression will be that in comparison with the norm group the school is gaining. But on a real NRT, their scores would go down dramatically because they were not being taught decimals at all.

Another sort of validity has to do with the difficulty of the sampled items. If the full range of difficulty is not included, the test artificially cuts off the bottom or top students, or maybe it jumps from very easy to very hard, and there is no discrimination among students.

There is a problem with test security both before and during the test. When stakes are high, when everyone is looking at test outcomes, there will be problems in some percentage of cases.

*"If we have a nonrandom sample of content... if the full range of difficulty is not included...(items) are then not norm valid.*

*"When stakes are high...there will be problems in some percentage of cases."*

# *Statewide Testing in Texas* -- *Test Publisher Perspectives*

...continued

***Paul Williams***

In some states a building test coordinator is made legally responsible for the security of tests. In Texas, the tests are delivered sealed, and if the booklets have problems (misprints, missing pages, etc.) it is a surprise for the teacher as well as the students. Some areas are trying to get staff other than teaching staff to administer the tests. So that they can say, "The teachers teach, then we test the students."

Another issue is teaching the test. When teachers, schools, districts are ranked on one number -- the percent who passed, or any other number -- the immediate reaction is to try to raise the number for its own sake. If the teacher's success is determined by how well his or her students do on a test, the test determines the curriculum. If the test is narrow, measuring only minimum skills, the teacher will drill forever on those minimum skills, and the test becomes the lowest common denominator on which all plans are based. If yes/no answers are required, the teacher who asks students to think and qualify and judge may not succeed as well as the teacher who just drills.

The choice is teaching to increase test scores versus teaching to increase learning across wide curricular areas. Teachers can all become familiar with tests given every year. We want scores to improve. We want to convince the Board, the newspapers, the taxpayers that we are doing it right. So the test determines the curriculum rather than in a true CRT where the curriculum determines the test.

*"...the immediate reaction is to try to raise the number for its own sake. ...the teacher will drill forever on those minimum skills..."*

*"The choice is teaching to increase test scores versus teaching to increase learning across wide curricular areas."*

# *Statewide Testing in Texas*

***-- Test Author Perspectives***

***H. D. Hoover***
***University of Iowa***

**The Regents Program** in New York may have been the first state testing program. Iowa uses the Iowa Tests of Basic Skills (ITBS) written at the University of Iowa, which owns the copyright. Riverside publishes the ITBS. Iowa buys the ITBS from Riverside to use in the state testing program.

Sometime around 1928, superintendents near Iowa City were tired of all the emphasis on sports in the spring. They talked to the Dean of Education at the University of Iowa about starting a spring academic contest. The test got very favorable publicity. It was well liked by people in the state, including legislators and the press. It started as an end-of-the-year, course-oriented, criterion-referenced test. After three or four years, it was obvious that it was a cheating disaster. It evolved into the Iowa Basic Skills Program in the mid-thirties. It has been a very stable program. I am only its third director in 54 years. When the ITBS was introduced in the elementary grades, the emphasis of the test changed to using test scores to improve instruction for individual students. There was a move away from the accountability aspects. The program still runs in much the same way.

In 1942, the program changed the high school test to fall administration only. Across the state, about 55 percent of the students are tested in the fall, 25-30 percent are tested at midyear, and the remainder are tested in the spring. The Iowa program is not mandated or controlled by the state. It is voluntary, but participation in it across the state is excellent. The university does not the want the program to be mandated, because they would lose control over the way the test results are used. There is a contract between the University of Iowa and each district. Results belong to the districts and are not given to the media by the university. Information on statewide trends can be released to the media. Each district gets data on statewide achievement trends.

Statewide testing tends to be political in most cases. What happens in Texas will not last more than ten years. What happens in statewide testing next in Texas will not be comparable to what happened before. These programs do not have much longevity and the uses of the results are focused in all the wrong ways. This occurs because the state legislature runs the testing program. Arthur Wise of the Rand Corporation wrote an article for the Kappan magazine in 1977 or 1978 called "Legislated Learning." In 1987, he wrote another article for the Kappan called "Legislated Learning--Revisited." Both articles are recommended reading about the outcomes of legislated state testing programs.

*"There is a contract between the University of Iowa and each district. Results belong to the districts and are not given to the media by the university."*

# *Statewide Testing in Texas*

***-- Test Author Perspectives***

...continued

***H. D. Hoover***

Test Security Issues —

When high stakes and high accountability are involved, test security is a problem. The only tests that are secure are tests such as the SAT and ACT where they spend $15 per student to ensure the security of the test. Achievement tests like the ITBS should not be secure. There is a new form of the ITBS which is supposed to be more secure.

Skills analysis summaries of achievement test results are provided for individual classes. If the purpose of testing is to improve instruction, teachers should have access to the tests to see where students had problems.

Quality of items can be a problem. The biggest problem of the proliferation of testing programs where everyone has their own CRT, or needs enough NRT items to sample from, is that there is not an infinite supply of good items. Tests must be reviewed by different groups to assure that items are not biased against certain groups. There is an infinite supply of bad test items. A major reason that some tests are so secure is so that no one can see how bad the items are.

Having equivalent forms has never been a problem in Iowa. Districts alternate between forms G and H each year. The test booklets are kept in Iowa City and are shipped to each district for testing. Booklets are shipped to principals if someone on the campus needs to look at them.

If the idea is to give out scores we have faith in for individual students, what use are customized programs? Customized programs don't go anywhere, but they do give states and the powers within the state the power of ownership.

A good sign is the trend back to individual scores for students. There seems to be a refocus in Texas on individual scores for individual students, but the emphasis will remain on comparing campuses and districts.

*"When high stakes and high accountability are involved, test security is a problem. If the purpose of testing is to improve instruction, teachers should have access to the tests to see where students had problems."*

*"...there is **not** an infinite supply of good items.*

*There **is** an infinite supply of **bad** test items."*

*"There seems to be a refocus in Texas on individual scores for individual students, but the emphasis will remain on comparing campuses and districts."*

# *Statewide Testing in Texas*

*-- Governmental Perspectives*

***Elaine Davis***
*Governor's Office*

**I am speaking today** as Elaine Davis who spent 20 years as a teacher and an administrator, and only six months in the Governor's Office.

I grew up in Spring Branch, where there was so much testing that by grade three some students were tracked and never seen again.

I was hired for two reasons:

1) Legislators want to know what their ideas would do to kids in public schools.

2) Information about school systems is power!!

I did a study of policy utilization of statewide testing data as a graduate student, and I found that the real estate people of Texas had data that school people could not get. Again, information is power.

Testing is important in Texas. Fifty percent of the State money is going into education. The Governor has an education team to support his belief that education is the bottom line of economic development. There is concern among these people about future uses of test data. However, in the future test scores will be used to determine where incentives go as well as for accountability.

Our economy now requires that 75% of our high school graduates know what only 25% knew in the 60's. We will either have to expand our school year or become more efficient. It takes lots of money to expand the school year or to make the current system more efficient, but a combination is coming. We need test data to see where the system is "doing it right" and to see where we are needing help. How can you make changes if you do not know where you are?

There is a myth that statewide testing was not intended for accountability. TEAMS was intended for accountability. We will always measure one class, school, district against another, based upon test scores. How can schools make changes if they do not know where they stand in comparison with other schools?

*"I grew up in Spring Branch, where there was so much testing that by grade three some students were tracked and never seen again."*

*"...in the future test scores will be used to determine where incentives go as well as for accountability."*

*"Our economy now requires that 75% of our high school graduates know what only 25% knew in the 60's."*

# *Statewide Testing in Texas* -- *Governmental Perspectives*

...continued

***Elaine Davis***

I would like to see a lot of control stay with educators rather than go to legislators.

In the past an educator could say, "I can do good things for children if you give me money." Now a person must say "I can prove it to you," or "this is the most cost effective way." Information is power and time is money. You need a one-page summary for busy legislators.

Educators need to look at business. Know "desk audits." Know and use business terminology to have an effect with the legislators.

Ask ourselves, "Are we who support school districts giving them the help they need?"

*"There is a myth that statewide testing was not intended for accountability."*

*"Are we who support school districts giving them the help they need?"*

# *Statewide Testing in Texas*

***SUMMARY STATEMENTS***

Whit Johnstone: Who's being held accountable? How? Qualitative data should also be used as part of this accountability?

Evangelina Mangino: In this whole testing effort, we need to remember to use the data to benefit students. The three most important issues in a statewide testing program are money, time, and usefulness of data.

Carl Shaw: There needs to be more concern with the sample of items used for the norm-reference scores on the next statewide test.

H.D. Hoover: Statewide testing allows legislators to appear concerned with education rather than spend much money. It is political. There is a problem with using business models in education, because we are not creating shavers that have to be identical. We are creating different people. Iowa is tops in achievement, but only midrange in the 50 states in terms of money spent. Iowa does have a concern with the education of its people.

Paul Williams: I am concerned with content spread, item difficulty, customized testing, and problems with percentiles with only a 20-item test. There is a need for interlevel articulation, or there may be a loss of floors and ceilings. We need to quit monkeying with shelf tests.

Mikel Brightman: Caution - we need to articulate why we are doing what we are doing, and what we are going to do with the testing program. We are obligated as professionals to clarify this. We should experiment with financial rewards for increasing test scores.

Elaine Davis: Legislators want 30 minutes or a one-page summary. The press wants a one-line headline. However, these are complex issues.

*Glynn Ligon: Question -- Are Texas schools more accountable and the students better educated after eight years of statewide testing?*

> Whit Johnstone: Yes, on accountability. No, to better educated.
>
> Carl Shaw: No, to better educated.
>
> Evangelina Mangino: No, to better educated for the higher achieving students. Yes, for the lower achieving students.

*"The three most important issues in a statewide testing program are money, time, and usefulness of data."*

*E.M.*